The impact of red noise in radial velocity planet searches: Only three planets orbiting GJ581?

Roman V. Baluev*

Central (Pulkovo) Astronomical Observatory of Russian Academy of Sciences, Pulkovskoje sh. 65/1, St Petersburg 196140, Russia
Sobolev Astronomical Institute, St Petersburg State University, Universitetskij prospekt 28, Petrodvorets, St Petersburg 198504, Russia

ABSTRACT

We perform a detailed analysis of the latest HARPS and Keck radial velocity data for the planet-hosting red dwarf GJ581, which attracted a lot of attention in recent time. We show that these data contain important correlated noise component ("red noise") with the correlation timescale of the order of 10 days. This red noise imposes a lot of misleading effects while we work in the traditional white-noise model. To eliminate these misleading effects, we propose a maximum-likelihood algorithm equipped by an extended model of the noise structure. We treat the red noise as a Gaussian random process with exponentially decaying correlation function.

Using this method we prove that: (i) planets b and c do exist in this system, since they can be independently detected in the HARPS and Keck data, and regardless of the assumed noise models; (ii) planet e can also be confirmed independently by the both datasets, although to reveal it in the Keck data it is mandatory to take the red noise into account; (iii) the recently announced putative planets f and g are likely just illusions of the red noise; (iv) the reality of the planet candidate GJ581 d is questionable, because it cannot be detected from the Keck data, and its statistical significance in the HARPS data (as well as in the combined dataset) drops to a marginal level of $\sim 2\sigma$, when the red noise is taken into account.

Therefore, the current data for GJ581 really support existence of no more than four (or maybe even only three) orbiting exoplanets. The planet candidate GJ581 d requests serious observational verification.

Key words: planetary systems - stars: individual: GJ581 - techniques: radial velocities - methods: data analysis - methods: statistical - surveys

1 INTRODUCTION

The multi-planet extrasolar system hosted by the red dwarf GJ581 has attracted a lot of interest in the past few years. The concise history of planet detections for this system is as follows. The first planet b, having orbital period of 5.37 d and minimum mass of $\sim 15 M_{\oplus}$, was reported by Bonfils et al. (2005). The two subsequent super-Earths c (with the orbital period of 12.9 d and the minimum mass of $\sim 5 M_{\oplus}$) and d (with the originally reported orbital period of 82 d later corrected to 67 d and current minimum mass estimate of $6 M_{\oplus}$) were discovered by Udry et al. (2007). Further, Mayor et al. (2009) reported the detection of the smallest exoplanet known so far, GJ581 e, orbiting the host star each 3.15 d and having the minimum mass of only appoximately $2 M_{\oplus}$. All these discoveries were done on the basis of the

radial velocity data obtained with the famous HARPS spectrograph.

Later, the Keck planet-search team got involved. Vogt et al. (2010) performed an analysis of the combined HARPS and Keck measurements and claimed the detection of two more planets in the system, f and g, orbiting the host star each 433 d and 36.6 d (and having minimum masses of ~ 7 and $\sim 3 M_{\oplus}$). The last planet g is remarkable because it appears to reside in the middle of the predicted habitable zone for this star. However, the reality of these two planets represents a subject of serious debates in the recent time. Gregory (2011) remained uncertain about the existence of these planets, based on his very detailed Bayesian analysis of the joint (Mayor et al. 2009) and (Vogt et al. 2010) datasets. Tadeu dos Santos et al. (2012) basically agreed with this conclusion, finding from the same combined dataset that the detection confidence probabilties for these two planets are 96% for planet g and 98% for planet f. These values are too high to be just neglected, but they are simultaneously too

low to claim a robust detection. Forveille et al. (2011) claim in a recent preprint that newer HARPS data do not support the existence of any planets beyond the four-planet model. Finally, in a very recent paper (Vogt et al. 2012) the authors assert, on the basis of the HARPS data from (Forveille et al. 2011), that with the false-alarm probability of $\sim 4\%$ an extra 32-day planet should exist in this system, beyond the four-planet model.

Summarizing these investigations, we must admit that the reality of the last detected planets is rather controversial. This uncertainty probably comes from some mysterious interference between the HARPS and Keck data. Indeed, it follows from e.g. (Gregory 2011) and (Tadeu dos Santos et al. 2012) that Keck data alone do not allow to detect more than only *two* planets b and c: all other planets seem to fall beyond the detection power of this time series. Newest HARPS data alone allow for the robust detection of *four* planets (from b to e) and do not really support the existence of the planets f and g. However, some additional variations can be still detected when the both datasets are joined, and it is rather uncomfortable just to ignore them.

In this paper we present an attempt to find a solution of this mystery. Our main idea comes from our previous work (Baluev 2011), where we analysed available radial velocity (RV) data for another planet-hosting red dwarf GJ876, and found that these data contain significant correlated noise component, also called as "red noise". Traditional statistical methods assume that the measurement errors are statistically independent, implying that their frequency power spectrum is flat (thus the noise is "white"). As we have shown in (Baluev 2011), both HARPS and Keck radial velocity measurements of GJ876 demonstrate non-white power spectra with a clearly visible excess at longer periods, and this non-whiteness is statistically significant. We found that a similar picture is often seen in the periodograms of the GJ581 data (see, e.g., a lot of periodograms plotted by Tadeu dos Santos et al. 2012). All this motivated us to investigate how the red noise could affect the derived orbital configuration of this system.

The structure of the paper is as follows. First, in the Section 2 we discuss the common undesired effects that the red noise might impose, and demonstrate how it reveal itself in the GJ581 RV data. In Section 3 we present a maximum-likelihood algorithm that can perform a reduction of the red noise, based on its full modeling. In the Section 4 we perform a detailed analysis of the latest radial velocity data for GJ581 taken from (Vogt et al. 2010) and (Forveille et al. 2011). We show how in the particular case of GJ581 the red noise creates fake RV variations, as well as hides the true ones. We also give two best fitting orbital solutions for this system, that take the red noise into account. In Section 5, we discuss the reality of the putative fifth and sixth planets, based on our RV data analysis. In the conclusive section of the paper, we discuss what global consequences the red noise implies for the past exoplanetary data-analysis works and what it requests from the future ones.

2   RED NOISE AS A MISLEADING AGENT

The routinely used methods of astrostatistics are designed to deal with the data containing uncorrelated noise. Such
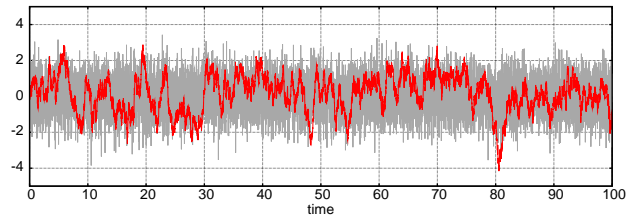

Figure 1. Foreground curve shows a simulated example of the red Gaussian noise with correlation function $e^{-|\tau|}$, while thick background band shows a simulated example of the white Gaussian noise of the same variance.

noise is also called white, because its frequency spectrum is flat: its periodograms demonstrate approximately the same mean level when averaged over different frequency segments of the same length.

In practice, however, the white-noise approximation may be poor. In particular, the noise in photometric observations of exoplanetary transits is routinely red (Pont et al. 2006). In the radial velocity planet searches, it is also known that the RV noise is not necessarily white, because it may demonstrate smaller level when averaged over larger timescales (e.g. O'Toole et al. 2008). However, for the RV case this issue basically appears as rather dark stuff with no routinely working practical solution known so far.

Potential impact of the red noise on the results of the data analysis may be huge. The correlated data usually carry smaller amount of information, as if their number was smaller. Therefore, when our data contain correlated noise, various statistical uncertainties are typically larger than we obtain based on the traditional white-noise models. It is the first effect imposed by the red noise. Another, possibly even more important effect, appears due to the non-uniform frequency spectrum of the red noise. Basically, the red noise is able to generate fake periodicities that can be mistakenly "detected" by the white-noise algorithms. This is demonstrated in Fig. 1, where the simulated example of a correlated noise looks like a bit noisy mixture of some illusive periodic signals. Moreover, these fake periodicities may obscure real variations, keeping them undetected until some data-analysis tool that is aware of the noise correlateness is applied.

In the case of GJ581, the white-noise model of the RV data is definitely inadequate. As we can see from Fig. 2, the periodograms of the residual noise that remains after elimination of the compound RV signal of 4 planets demonstrate clear excess of the power at low frequencies ($\lesssim 0.1$ day^{-1}) a symmetric excess around 1 day period (emerging due to a strong diurnal aliasing), and a depression in the middle of the segment. It is important that this power excess does not concentrate in any well-defined discrete peaks; instead it is spread smoothly in a continuous frequency band. The both periodograms for the HARPS and Keck data demonstrate a similar smoothed shape, although the positions of individual high peaks have little common (meaning that all relevant periodicities are not real). We may note that this picture looks very similar to the one we have already seen for the GJ876 case (Baluev 2011).

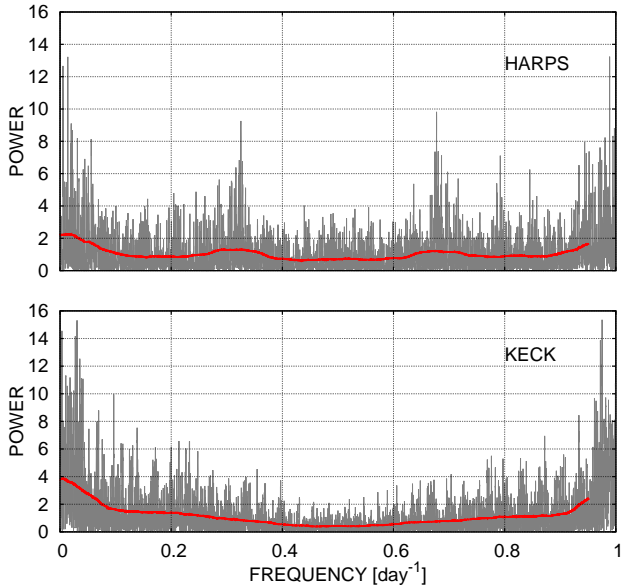The RV noise correlation can be also revealed in the

Figure 2. The red noise in the HARPS and KECK RV data for GJ581 in the frequency domain. We plot the residual periodograms that remain after elimination of the compound 4-planet signal. These periodograms are constructed separately for the HARPS and Keck data by means of adding the probe sinusoidal signal to the RV model of only HARPS or only Keck dataset, but still fitting such RV model jointly to the both datasets. Each value of these periodograms represents the modified likelihood ratio statistic $\tilde{Z}$ defined in (Baluev 2009). The smoothed curves represent the moving average of the raw periodograms obtained using the window of 0.1 day^{-1}.
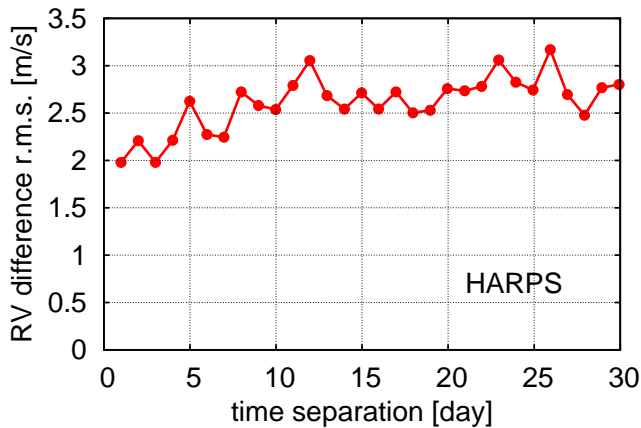

Figure 3. The red noise in the HARPS RV data for GJ581 in the time domain. See text for the detailed description.

time domain. This is easy to do for the HARPS data thanks to the following favouring factors:

(i) All HARPS measurements were done at almost the same sidereal time, implying that the time separation between two arbitrary datapoints is usually very close to an integer number of sidereal days.

(ii) There are a lot (up to ~ 100) of the HARPS RV pairs that have a small time separation of only one or a few sidereal days.

We run the following procedure. Given some integer n from 1 to 30, we collect all pairs of the HARPS measurements that are separated by n sidereal days. For each such pair we evaluate the difference of the 4-planet RV residuals corresponding to the observations involved in this pair. Finally, for each group of the RV pairs with a given time separation we evaluate the sample variance of this RV difference. How this variance can help us to detect red noise? Assuming that the variance of the residual RV noise has some constant value of σ^2 (which is not too far from the truth) and its autocorrelation function is $R(\Delta t)$, the variance of the mentioned above RV difference should be $2\sigma^2(1-R(\Delta t))$. Therefore, the graph of this variance should basically represent a rescaled upside-down view of the noise correlation function. We show this plot in Fig. 3. We can see clear growing trend before the time separation of 10 days, and a saturation beyond this point. Basically, the HARPS RV measurements have better relative precision at short timescales of up to ~ 10 days, while at longer time separations they show larger random scatter.

Unfortunately, the distribution of the Keck data points is not regular enough, and application of a similar procedure to the Keck dataset was not informative.

In this section we limit ourself by demonstration only, leaving the rigorous determination of the red noise significance for further sections. However, it is already clear that we may obtain trustable results concerning the GJ581 planetary system only if we utilize some method of the data analysis that can properly deal with correlated noise.

3 MAXIMUM-LIKELIHOOD REDUCTION OF THE RED NOISE

The method that we propose for the analysis of the data polluted by correlated noise represents a generalization of the maximum-likelihood approach described in (Baluev 2009), which was already used in (Baluev 2011). The main idea of the method is to construct some suitable correlational model of the noise in the RV data and then, based on this model and on the Keplerian model of the RV curve, apply the maximum-likelihood alorithm to estimate the parameters of the both models and the relevant goodness of the fit.

Thus, first we should choose some realistic and simultaneously simple model of such correlated noise. We assume that this noise is a Gaussian random process with some known correlation function. This means that the full vector of our N RV measurements x_i, taken at the timings t_i, should follow a multivariate Gaussian distribution. The mean of this distribution equal to the RV curve model $\mu(t_i, \boldsymbol{\theta})$, and the relevant variance-covariance matrix is $\mathbf{V}(\boldsymbol{p})$, where the vectors $\boldsymbol{\theta}$ and $\boldsymbol{p}$ contain some free parameters to be estimated from the data. The corresponding log-likelihood function may be expressed as

$$\ln \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{p}) = -\frac{1}{2} \ln \det \mathbf{V} - \frac{1}{2} \boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \boldsymbol{r} + N \ln \sqrt{2\pi}, \quad (1)$$

where $\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})$ is the vector of the RV residuals. For shortness, we will also denote the combined vector of all parameters $\boldsymbol{\theta}$ and $\boldsymbol{p}$ as $\boldsymbol{\xi}$.

Maximizing (1) over $\boldsymbol{\theta}$ and $\boldsymbol{p}$, we obtain the best fitting

4 R.V. Baluev

estimations of these parameters (the point where the maximum is attained). The value of the likelihood maximum itself may further be used, for instance, in the likelihood ratio test comparing two different data models.

As we explain in (Baluev 2009), in practice it might be useful to replace the true likelihood function (1) by a modified version

$$\ln \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{p}) = -\frac{1}{2} \ln \det \mathbf{V} - \frac{1}{2\gamma} \boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \boldsymbol{r} + N \ln \sqrt{2\pi}, \qquad (2)$$

where the correction divisor $\gamma = 1 - \dim \boldsymbol{\theta}/N$. The goal of this modification is to reduce the systematic bias that would otherwise appear in the noise parameters, because the best-fit residuals $\boldsymbol{r}$ are systematically smaller than real errors. This effect is clear, e.g., in Fig. C2 that we will discuss in detail in a further section. The bootstrap simulation in this plot (left panel) show clear systematic bias because the bootstrap is based on the unscaled best-fit RV residuals. If not the correction (2), the plain Monte Carlo simulations (right panel) would demonstrate the same or simular bias, while the bias of the bootstrap would be effectively doubled. Note that the modification (2) keeps intact all asymptotic properties of the maximum-likelihood method; it only improves its behaviour when N is not so large.

Note that the generalized model (1) differs from the one used in (Baluev 2009) in the matrix $\mathbf{V}$, which is no longer diagonal. However the general theory of maximum-likelihood estimations is basically the same for the both cases. For example, to find the covariance matrix of the maximum-likelihood estimations, we should first calculate the quadratic Taylor approximation of the function $\ln \mathcal{L}(\boldsymbol{\xi})$. From (1) we can easily derive the relevant gradient:

$$\begin{aligned}\frac{\partial \ln \mathcal{L}}{\partial p_i} &= -\frac{1}{2} \mathrm{Tr} \left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_i} \right) + \frac{1}{2} \boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_i} \mathbf{V}^{-1} \boldsymbol{r}, \\ \frac{\partial \ln \mathcal{L}}{\partial \theta_i} &= \boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_i}, \end{aligned} \qquad (3)$$

as well as the second-order derivatives:

$$\begin{aligned}\frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j} &= \frac{1}{2} \mathrm{Tr} \left[\mathbf{V}^{-1} \left(\frac{\partial \mathbf{V}}{\partial p_i} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_j} - \frac{\partial^2 \mathbf{V}}{\partial p_i \partial p_j} \right) \right] - \\ &\quad - \boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \left(\frac{\partial \mathbf{V}}{\partial p_i} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_j} - \frac{1}{2} \frac{\partial^2 \mathbf{V}}{\partial p_i \partial p_j} \right) \mathbf{V}^{-1} \boldsymbol{r}, \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} &= -\frac{\partial \boldsymbol{\mu}^{\mathrm{T}}}{\partial \theta_i} \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_j} + \boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \frac{\partial^2 \boldsymbol{\mu}}{\partial \theta_i \partial \theta_j}, \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial p_j} &= -\boldsymbol{r}^{\mathrm{T}} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_j} \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_i}. \end{aligned} \qquad (4)$$

Considering together (1-4), we can write down the following quadratic approximation:

$$\ln \mathcal{L}(\boldsymbol{\xi}) \simeq \ln \mathcal{L}(\hat{\boldsymbol{\xi}}) + \boldsymbol{g} \cdot \Delta \boldsymbol{\xi} - \frac{1}{2} \Delta \boldsymbol{\xi}^{\mathrm{T}} \mathbf{F} \Delta \boldsymbol{\xi}, \qquad (5)$$

where $\Delta \boldsymbol{\xi} = \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}$ with $\hat{\boldsymbol{\xi}}$ standing for the vector of the true parameters, $\boldsymbol{g} = \partial \ln \mathcal{L}/\partial \boldsymbol{\xi}$ is the compound gradient of $\ln \mathcal{L}$ and $\mathbf{F}$ is the Fisher information matrix:

$$\mathbf{F} = \mathbb{E} \left(\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\xi}} \otimes \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\xi}} \right) = -\mathbb{E} \frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\xi}^2}, \qquad (6)$$

where the expectation should be taken at the true parameter values. The elements of $\mathbf{F}$ in our case are

$$F_{p_i p_j} = \frac{1}{2} \mathrm{Tr} \left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_j} \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial p_i} \right),$$

$$\begin{aligned}F_{\theta_i \theta_j} &= \frac{\partial \boldsymbol{\mu}^{\mathrm{T}}}{\partial \theta_i} \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_j}, \\ F_{\theta_i p_j} &= 0. \end{aligned} \qquad (7)$$

The expansion (5) allows us to approximate the point $\boldsymbol{\xi}^*$, where the maximum is achieved, as $\boldsymbol{\xi}^* \simeq \hat{\boldsymbol{\xi}} + \mathbf{F}^{-1} \boldsymbol{g}$. Since the relation $\mathrm{Var} \, \boldsymbol{g} = \mathbf{F}$ holds true, the variance-covariane matrix of our estimations $\boldsymbol{\xi}^*$ has the same asymptotic representation for large N as in the uncorrelated case:

$$\mathrm{Var} \, \boldsymbol{\xi}^* \simeq \mathbf{F}^{-1}. \qquad (8)$$

Notice that the vectors $\boldsymbol{\theta}^*$ and $\boldsymbol{p}^*$ appear therefore asymptotically uncorrelated, as in (Baluev 2009), since the cross term $F_{\boldsymbol{\theta} \boldsymbol{p}}$ is again zero.

The numerical non-linear maximization of (1) or (2) can be performed by means of the Levenberg-Marquardt algorithm. Notice that the simplified widespread version of this algorithm that minimizes a sum-of-squares function (as implemented, e.g., in the MINPACK package) is unsuitable here, because it relies on certain relationships between the gradient and Hessian matrix, which are invalid in our case. We need a general variant of the Levenberg-Marquardt algorithm (e.g. Bard 1974) that can maximize an arbitrary non-linear target function and can deal separately with the gradient and the Hessian. Note that it is handy to approximate the Hessian as $-\mathbf{F}$ due to the expansion (5). It is useful because $\mathbf{F}$ is always positive definite. Besides, a lot of care is needed to optimize the calculational performance, since the formulae (3)-(7) require very computation-greedy operations with large matrices and vectors. We give some tips concerning this issue in the Appendix A.

We only need to detail the last thing, namely the model of the noise covariance matrix $\mathbf{V}$. In this paper we consider three main noise models:

(i) White-noise model. The matrix $\mathbf{V}$ is diagonal; the diagonal elements represent the total variances of individual RV measurements and are equal to the sum of the instrumental part (the square of the stated measurement uncertainty) and the RV jitter. The RV jitter is different for different instruments. This model was considered in (Baluev 2009).

(ii) Shared red-noise model. In addition to the white-noise components, we add to $\mathbf{V}$ the red-noise covariance matrix $\sigma_{\mathrm{red}}^2 \mathbf{R}(\tau)$, where the elements of $\mathbf{R}$ are defined via some guessed noise correlation function $\rho(x)$ as $R_{ij}(\tau) = \rho((t_j - t_i)/\tau)$. We chose $\rho(x) = e^{-|x|}$. This noise model infers that the red noise belongs to the star, while the spectrographs generate only the white noise.

(iii) Separated red-noise model. It is similar to the previous case, but the parameters σ_{red} and τ are different for different instruments (HARPS and Keck). The cross correlation between HARPS and Keck measurements is set to zero. This model infers that the red noise belongs to the spectrographs, and not to the star.

4 GJ581 DATA ANALYSIS

4.1 Preliminary investigation

The main goal of this subsection is to estimate the validity of various statistical methods in the case of GJ581 RV

data analysis. We have two datasets at our disposal: the 240 HARPS and 121 Keck/HIRES RV measurements published in (Forveille et al. 2011) and (Vogt et al. 2010), respectively. First of all, we provide four-planet white-noise fit in Table 1 that was obtained by means of the likelihood function maximuzation as described in (Baluev 2008b, 2009, 2011).

The maximum-likelihood approach infers a set of well-known classical theoretical results and methods concerning the maximum-likelihood estimations, that were established under a condition that the number of observations tends to infinity. We will call them collectively as Asymptotic Maximum-Likelihood Estimation Theory (hereafter AMLET). A fraction of them is described in (Baluev 2009), bearing in mind an application to the exoplanetary RV curve fitting task.

Notice that AMLET tools are sometimes called as frequentist ones, especially in the works employing the Bayesian analysis, where such opposing highlights the advantages of the Bayesian methods. Such terminology actually hides a misconception: AMLET represents, basically, a common limit to which both frequentist and Bayesian approaches converge, when the number of observation tends to infinity. Therefore, it would be incorrect to equate AMLET and the general frequentist approach in the statistics, since most of the classical AMLET propositions can be easily reinterpreted from the Bayesian point of view, while the genuine frequentist methods in their general form (e.g. Lehman 1959) are more complicated and theoretically justified than AMLET.

It is often argued that AMLET tools are often not applicable to the exoplanetary RV data analysis, especially for multi-planet systems which involve very complicated non-linear RV signal models. However, this is rarely verified with concrete practical cases. In this paper we undertook an attempt to rigorously assess the applicability of AMLET tools to the case of the GJ581. Since this research invloves a vast amount of various numerical simulations and rather boring statistical stuff, which is not related directly to the GJ581 system itself, we do not describe these results in the main body of the paper. An interested reader may find the detailed discussion in the Appendix C. Here we only provide a short summary of our investigation:

(i) In the case of the GJ581 RV data, the parametric confidence regions and false alarm probabilities, obtained using AMLET, work well for the white-noise and shared red-noise 4-planet models, but are unsuitable for the separated red-noise model. This indicates that the latter model is overparametrized and must be used with caution.

(ii) When analysing the HARPS and Keck data independently from each other, we may use AMLET for the HARPS time series, but not for the Keck one. Actually, the Keck dataset is the main thing that makes AMLET unusable with the separated red-noise RV model.

(iii) We should avoid using the bootstrap simulation (section B2) for any of our models, because it works in an unexpected and misleading manner when a parameterized noise is involved. However, we may safely use the usual Monte Carlo simulation (section B1) instead, since the RV data show absolutely no hints of any non-Gaussianity, which is the main fear of people preferring the bootstrap.[1]

(iv) In the most cases, we have no need for complicated and computationally greedy techniques like the Bayesian analysis or the genuine frequentist methods. For the white-noise and shared red-noise model we can just use AMLET with no fear. For the separated red-noise model AMLET is poor, but with this model we obtain little serious results that would need a deep verification.

4.2 Assessing the significance of the red noise

Let us first assess rigorously the significance of the noise non-whiteness. We can do this using the method described in (Baluev 2011). Using a variation of the Monte Carlo algorithm B1 from the Appendix B, we generated a bunch of 1000 simulated residual periodograms assuming the white model of the noise and 4-planet model of the RV curve. Thus, these periodograms are evaluated using exactly the same algorithm as in Fig. 2, but based on simulated uncorrelated data. Each simulated periodogram is then smoothed, also exactly as in Fig. 2. Based on this set we can derive the distribution of single values of the smoothed periodograms (for an arbitrary frequency), and also the distribution of the associated max / min ratio, which characterizes the degree of non-whiteness of the simulated spectrum. It is important that this procedure does not require us to make any assumptions about the red-noise correlation function.

The results are shown in Fig. 4. We can see that among 1000 Monte Carlo trials none could reproduce the same large max / min ratios that we observed for the smoothed periodogram of the real data. Therefore, the non-whiteness in the RV noise of these real data has very high significance ($> 99.9\%$). For comparison, we also plot the periodograms of the real data on the basis of the red-noise models, utilising the algorithm of Section 3. We can see that these frequency spectra are already consistent with the white-noise statistical limits, possibly except for the case of the Keck periodogram with shared red-noise model, where the residual non-whiteness has the significance of 2.6σ. Therefore, this model may be incapable of complete elimination of the red noise, probably because the red noise has somewhat different characteristics between the HARPS and Keck datasets. We think, however, that this shared red-noise model suits our practical needs at best, since the residual frequency spectrum non-whiteness is anyway a few times smaller than it was for the original white-noise model. The separated red-noise noise model can do apparently more impressive reduction of the correlated noise, but, as we have discussed above, this model is statistically poor.

We must emphasize that all significance levels printed in Fig. 4, including the one of 2.6σ for the Keck shared red-noise periodogram, were derived from white-noise simulations. To be fully honest, we ought to evaluate these levels assuming a matching noise model for each, but it

---

[1] Sometimes it is claimed that a correlated noise is not Gaussian. We must caution the reader against such mixing of distinct notions. In our case of GJ581, for instance, the noise is consistent with a Gaussian random process. Such a process has Gaussian individual values, which are nonetheless mutually correlated.

Table 1. Best fitting parameters of the GJ581 planetary system: white jitter, four planets

	planetary parameters			
	planet b	planet c	planet d	planet e
P [day]	5.368585(79)	12.9175(19)	66.616(79)	3.14922(18)
$\tilde{K} = K\sqrt{1-e^2}$ [m/s]	12.58(16)	3.26(16)	1.95(17)	1.79(16)
e	0.021(13)	0.053(48)	0.259(83)	0.164(89)
ω [°]	334(35)	145(53)	342(18)	156(31)
λ [°]	142.93(76)	106.5(3.0)	144.9(5.2)	63.3(5.4)
$M\sin i$ [$M_\oplus$]	15.78(20)	5.48(27)	5.65(49)	1.88(17)
a [AU]	0.04061187(40)	0.0729244(70)	0.21768(17)	0.0284573(11)
	data series and common fit parameters			
	HARPS	Keck		
c_0 [m/s]	−9205.96(13)	1.08(27)		
σ_{white} [m/s]	1.50(11)	2.45(23)		
r.m.s. [m/s]	1.96	2.82		
$\tilde{l}$ [m/s]		2.25		

The $M\sin i$ and a values were derived assuming the mass of the star $M_\star = 0.31 M_\odot$, which was used e.g. by Forveille et al. (2011). The uncertainty of $M_\star$ were not included in the uncertainties of the derived values. The mean longitudes λ refer to the epoch $JD2454500$. The goodness of the fit $\tilde{l}$ was derived from the maximum of $\ln\mathcal{L}$ as explained in (Baluev 2009).
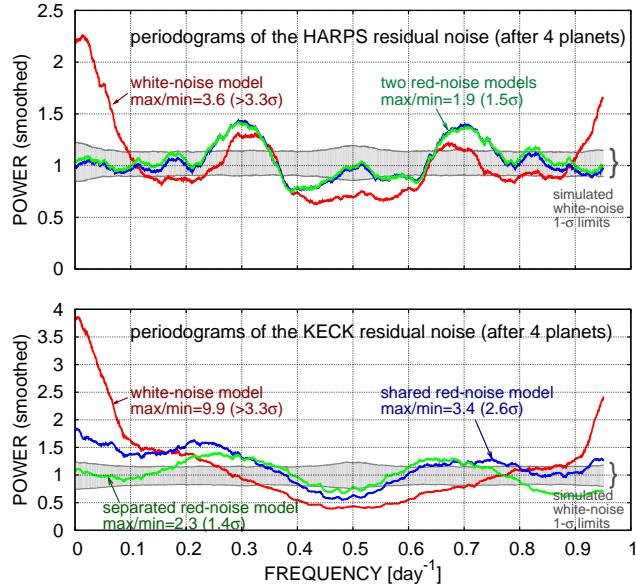

Figure 4. The periodograms of the HARPS and Keck residual noise constructed for various noise models, in comparison with simulated limits expected for the white-noise case. We can see that periodograms based on the white-noise model show large variations with max / min ratio well above the 3σ level, and using the red-noise models significantly suppresses these variations.

appeared too computationally-demanding for the red-noise cases. We expect that the correct significances for red-noise periodograms may be somewhat smaller, because such periodograms usually showed systematically higher significance levels. It may even appear, that the mentioned residual Keck non-whiteness for the shared red-noise case is eventually insignificant. However, this does not alter our conclusion that the white-noise model is inadequate; the relevant significance is based on the correct (matching) noise model and

is well above the 3σ level, both for the HARPS and Keck data.

4.3 Detailed analysis of the RV data

4.3.1 HARPS data alone

Let us start from the analysis of 240 HARPS RV measurements. In Fig. 5 we show a series of the residual periodograms, starting from the two-planet base model (planets b and c). In the case of the white-noise model we are able to subsequently extract all four planets from these periodograms. We can see that all four peaks show very high significance. However, in the last residual periodogram, corresponding to the case when all four peaks are already extracted, we can see a typical red noise picture: an amorphous set of peaks at the periods longer than ~ 10 d, a diurnal alias of this frequency band close to 1 d period, and a depression in the middle part of the period range.

We can see that our maximum-likelihood algorithm suppressed the effect of the red noise, as expected. However, together with the red noise, our procedure dramatically suppressed the planet d peak at 67 d. This is not very surprising on itself, since this peak is in the range where the red noise is ruling. However, the final significance of this peak becomes marginal – only 1.8σ – making us rather sceptical about the reality of this planet.

Speaking shortly, although we cannot claim that the planet d RV signature is insignificant, we must admit that its detection is not robust and requires a serious verification. The relevant RV variation may be caused by correlated RV noise in the data, and does not necessarily reflect a Doppler wobble induced by a real planet.

4.3.2 Keck data alone

Let us now deal with 121 Keck/HIRES RV measurements in the similar manner. The relevant periodograms are shown

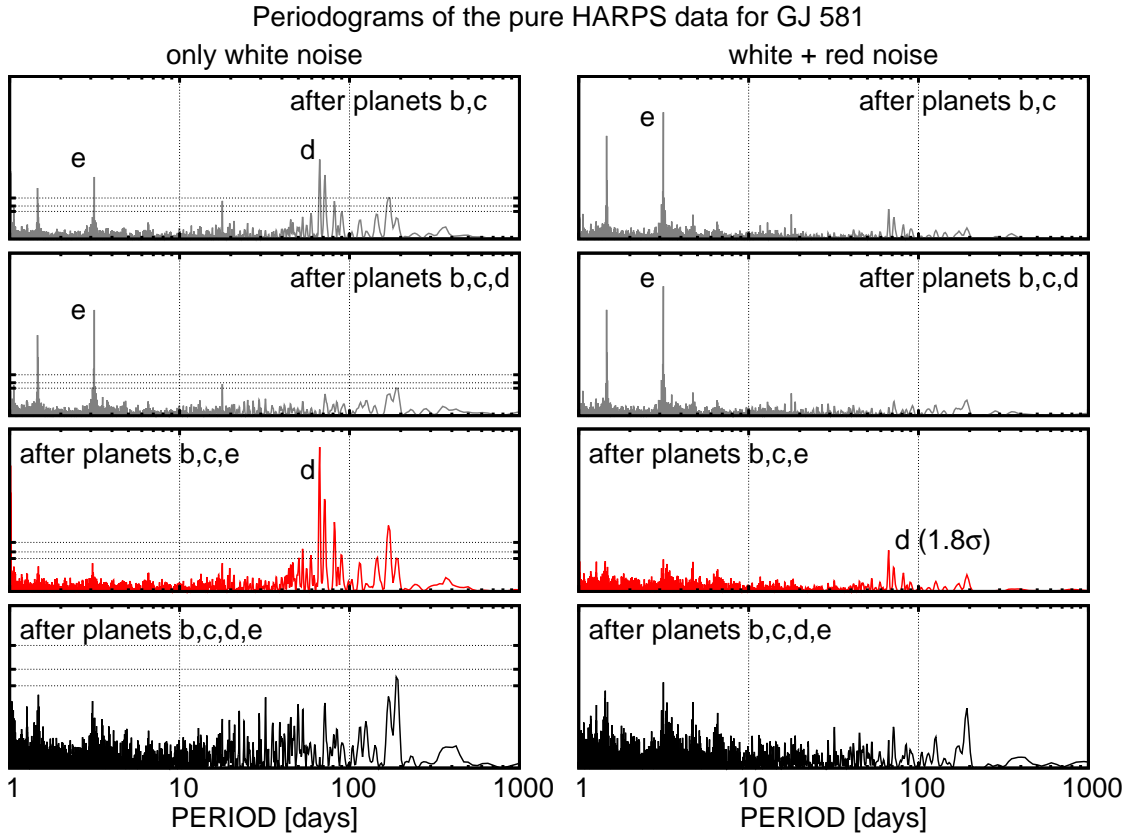Periodograms of the pure HARPS data for GJ 581

Figure 5. Residual periodograms of the HARPS RV data of GJ581 constructed on the basis of various base RV curve models. The base models include the compound multi-Keplerian signal from the planets labelled in a relevant plot. Note that the plots in the same row always have the same ordinate range. In the white-noise plots we show the 1-, 2-, and 3-σ significance levels derived in accordance with (Baluev 2008a). These levels are in good agreement with Monte Carlo simulations. The technique from (Baluev 2008a) was not designed to work with correlated data, so for the red-noise plots we only show the simulated significance of the most interesting planet d peak (third periodogram in the red-noise column).

in Fig. 6. First, we can see that now we cannot detect more than two planets b and c, if we use the traditional white-noise model. This conclusion is in the agreement with previous studies (Gregory 2011; Tadeu dos Santos et al. 2012), but it is still rather disappointing, because the Keck RV precision is pretty competitive in comparison with the HARPS one. Second, in the Keck data we again see clear hints of the red noise, which created a fake periodicity at approximately 27 days. In this case our red-noise removing algorithm does its job even better than anyone could expect: it did not just killed all fake red-noise peaks, but it also reveals the 3.1-day peak belonging to the planet e! This proves that our algorithm does not just suppresses the apparent RV variations, lifting up the detection thresholds. It is working in a much more intelligent manner: in certain frequency ranges it may basically *improve* the effective RV precision, revealing the true periodicities that the red noise tries to hide.

    The period of the newly discovered variation in the Keck data is in excellent agreement with the planet e period obtained from the HARPS data. Such coincidence is hardly casual. However, what is its rigorous significance? The answer to this question is not obvious, because we cannot use AMLET for the Keck dataset alone. Monte Carlo simulations (algorithm B1) suggest that the significance associated to

this peak of the Keck periodogram is only 1.2σ. Therefore, if we tried to *detect* this planet from the Keck data alone, with absolutely no reference to the HARPS data, we would have to admit that this peak is statistically insignificant. Basically, it is a luck that no other comparable peak appeared in the top-right periodogram of Fig. 6.

    However, we need just to *confirm* the planet e existence on the basis of the Keck dataset, not to *detect* it anew. This places much more mild limits. We have no need to simulate the periodogram in its whole period range as shown in Fig. 6. We already know the probable planet e parameters from the HARPS data with good precision, including e.g. its orbital period. Now we only need to confirm that RV noise could not generate so large peak as we can see in the Keck data just *in a narrow vicinity* around this known period. It is not a big deal if we find a noisy peak at some faraway frequency, where the real planet e definitely cannot reside. Such *confirmational* significance will be much larger than the *detectional* one, because the probability for the RV noise to occasionally generate a large peak inside a narrow frequency segment is much smaller than inside a wide one. This becomes obvious if we look at the periodogram's false-alarm probability approximation from (Baluev 2008a):

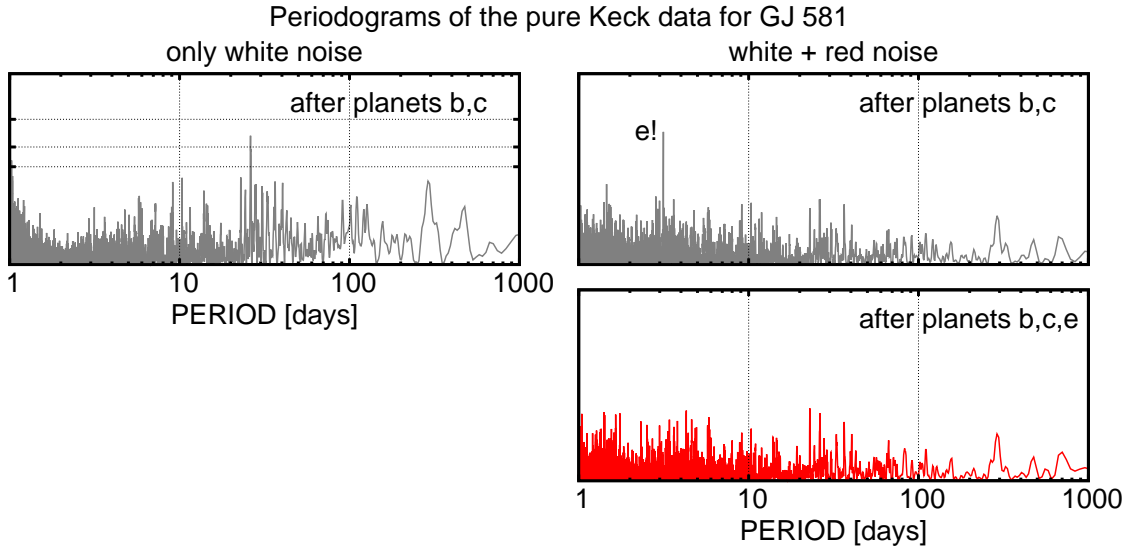$$\text{FAP} \lesssim W e^{-z} \sqrt{z}. \tag{9}$$

Figure 6. Same as in Fig. 5, but for the Keck dataset. We can see that the white-noise model does not allow robust detection of any real planet except for the planets *b* and *c* (which both are undoubtful anyway), but applying the red-noise model allows to confirm the existence of the planet *e* (see text for the detailed calculation of its significance). No hints of the planet *d* can be revealed with any of these models, however.

We can see that this estimation depends on the normalized frequency bandwidth W. For the whole range of periods from 1 day to infinity we have $W \approx 3500$ (for the Keck dataset taken alone), but when dealing with confirmational false alarm probabilities, we need to consider $W \sim 1$ at most, since the $\pm 1\sigma$ uncertainty range of the expected planet *e* period corresponds to $W \sim 0.1$. Therefore, the confirmational false alarm probability might be a few thousand times smaller that the detectional one.

However, we must remember that the red-noise model infers strong non-linearity when used with Keck data alone, as we have already discussed above. We cannot use any asymptotic methods in this case, and our numerical simulations must be more intricate than the plain Monte Carlo scheme B1. We can no longer rely on a single simulation series based on a single vector of nominal "true" model parameters, as we have done before. Instead, we should honour some representative parametric domain. Since the true values may be anywhere in this domain, we must generate many distinct simulation series according to the Monte Carlo scheme from Section B1, each time assuming different vector for the mock "true" parameters. The detailed step-by-step description of this algorithm is given in Section B3.

During this calculation, we first generated a sequence of the trial "true" sets of parameters, assuming the two-planet red-noise Keck RV data model (the first-level simulation). This first-level simulation is not intended to have big statistical meaning, we need it just to obtain a set of points covering some more or less wide domain around the best fitting two-planet solution. After that, for each of the generated parametric vectors, we run the plain Monte Carlo simulation (algorithm B1, 500000 random trials in each simulation). On each random trial of this second-level simulation we generate an artificial Keck dataset using the model "two-planet RV variation + correlated RV noise" (without planet *e*). Then for each such dataset we evaluate the rel-

evant residual periodogram exactly in the same way as in the top-right panel of Fig. 6, where we used the real Keck data. For each such periodogram we find the maximum in the narrow period range $3.145 - 3.153$ day ($W \approx 3$, centered at the nominal planet *e* period). Based on such Monte Carlo sequence, we count the fraction of simulated periodograms that demonstrated the same or larger maximum peak as the one that we have seen for the real data. This fraction represents the desired confirmational false-alarm probability of the planet *e*, as inferred by the adopted "true" two-planet model. These false-alarm probabilities can be further transformed to the normal ("n-σ") significance levels that we use throughout the paper.

We plot the results of these simulations in Fig. 7. From this graph, we can see that the simulated confirmational significance practically does not depend on the adopted parameters of the base two-planet model, even when these parameters deviate from the nominal estimation by more than 2σ. Actually, most of the scatter around the nominal level is likely due to the statistical uncertainty of the second-level Monte Carlo. If we generated more trials for each point in Fig. 7, this scatter would probably shrink further. Basically, this figure does not reveal any real dependence on the true parameters (at least in the parametric domain that we were able to fill in the first-level Monte Carlo). The rigorous frequentist significance level is given by the minimum ordinate among all simulated points, while the nominal level corresponds to the point located at zero abscissa. These values do not differ much and can be rounded to 3σ both. This means that Keck data can robustly confirm the existence of the planet GJ581 *e* RV signal.

Our algorithm does not reveal any hint of the planet *d* in the Keck data. The corresponding residual periodogram calculated after extraction of the three planets *b*, *c*, and *e*, looks like a perfect white noise with no peak attracting any attention. Maybe this planet *d* does not actually exist,
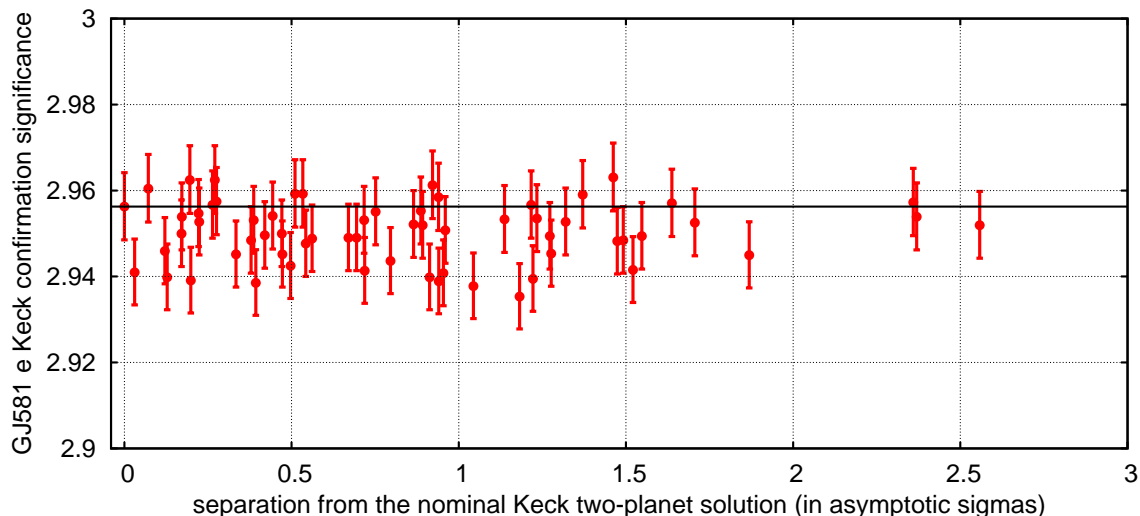
Figure 7. Simulated confirmation significance of the planet GJ581 *e*, depending on the assumed true parameters of the two-planet red-noise model of the Keck RV data. The graph shows a set of points, each corresponding to a different test "true" vector of parameters, simulated according to the Monte Carlo algorithm B1. The abscissas of the points reflect the deviation of these trial "true" parameters from the nominal best-fit ones, expressed in terms of the asymptotic n-σ significance level, which was derived from the relevant value of $\tilde{Z}$. The ordinate of each point shows the confirmation significance of the planet e, as derived from 500000 Monte Carlo simulations of the relevant Keck periodogram (see text for details). The errorbars reflect the statistical uncertainties inferred by these second-level Monte Carlo simulations. We can see that the confirmation significance almost reaches the 3σ level and is practically independent on the assumed two-planet configuration.

and the variation that we have seen in the HARPS data is some systematic effect or just some random fluctuation? Unfortunately, Keck data alone cannot supply an independent answer to this question. Notice that there is a pretty large difference in the significance of the planet e, as inferred by the HARPS and Keck data. After extrapolation of this difference to the planet d, we realize that currently available Keck data are just unabe to reveal it, no matter exists it or not.

4.3.3 *Combined dataset*

Now let us proceed to the joint analysis of the HARPS and Keck data. We consider three noise models that we have already introduced: the white-noise model, the shared red-noise model, and the separated red-noise model.
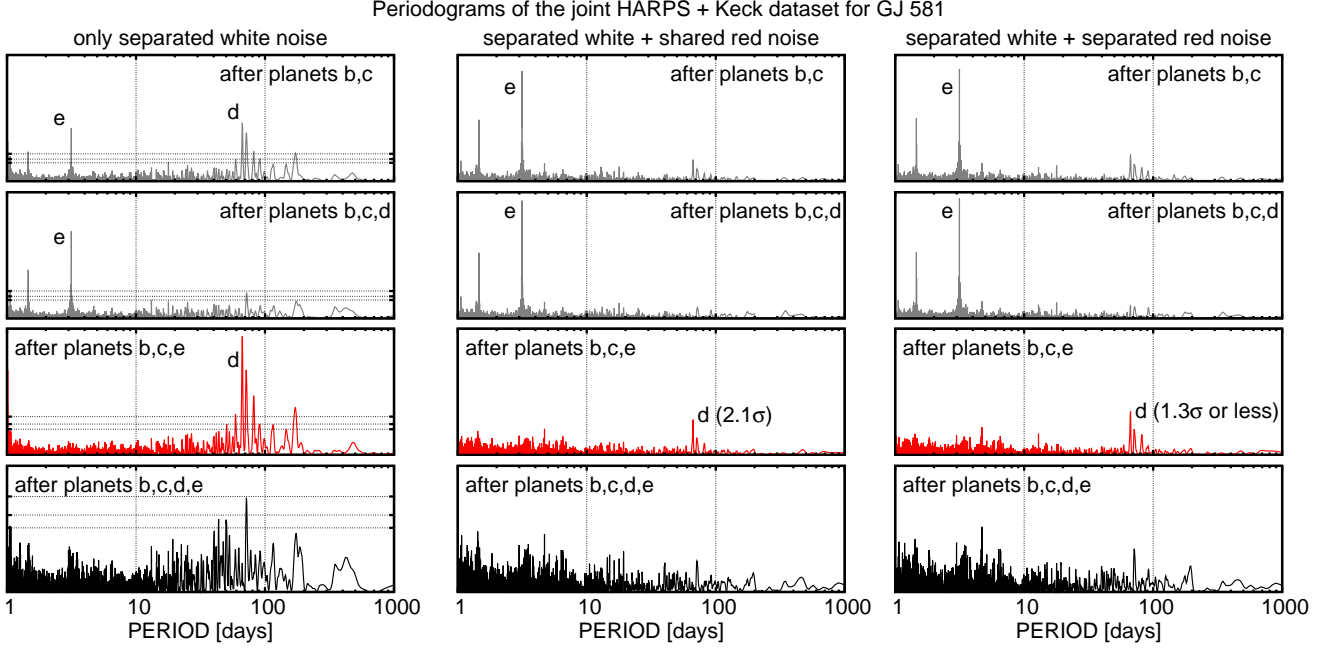
We show a series of the relevant periodograms in Fig. 8. We can see that while the planets b, c, and e can be robustly extracted from these data, the planet d still remains rather controversial, because its significance drops to only $\sim 2.1\sigma$ or even below, if a red-noise model is used. We feel such significance level is not enough to claim a robust detection of an exoplanet, because this significance is model-dependent. The planet candidate GJ581 d should be probably reclassified as a controversial one.

Finally, we present two fits of the GJ581 planetary system, obtained using the shared red-noise model. In the first fit (Table 2) we provide a three-planet configuration with planets b, c, and e, while the second one (Table 3) also involves planet d.

5   REALITY OF THE PUTATIVE FIFTH AND SIXTH PLANETS

Various authors have already raised a lot of doubts concerning the existence of the planets GJ581 f and GJ581 g, since their announcement in (Vogt et al. 2010). Basically, it appears that their parameters may change significantly with each data update and, besides, they do not demonstrate enough resistance with respect to various rather subjective choices: methods of data analysis, model details, etc. For example, even for the traditional white-noise model, we do not see in our work any hints of these planets, as they were originally reported in (Vogt et al. 2010). The four-planet residual periodograms plotted here separately for the HARPS, Keck, and joint datasets demonstrate different patterns of individual marginally significant maximum peaks, although they all reveal a similar average power excess at the periods longer than 10 days.

Our advanced analysis suggests that the red-noise models leave no room for any RV variations beyond the four-planet models: neither in the HARPS, nor in the Keck, nor in the combined dataset. Our red-noise models just absorbed all RV signals interpreted previously as the hints of the planets f and g. However, maybe it is just a question of interpretation? There are a lot of resons explaining why the measuments taken at different epoch may be statistically correlated with each other. The genuine RV noise caused by some astrophysical activity of the star, for instance, is a likely explanation, but other sources are still possible. For example, a sinusoidal periodic oscillation would produce the same periodic contribution in the compound correlation function of the data. This might produce a picture very similar to what we see e.g. in Fig. 3 until we extend this graph to larger time lags (which appears practically impossible because the

Figure 8. Same as in Fig. 5, but for the combined HARPS+Keck dataset.

Table 2. Best fitting parameters of the GJ581 planetary system: shared red jitter, three planets

	planet b	planet c	planet e
planetary parameters			
P [day]	5.368589(68)	12.9186(21)	3.14905(16)
$\tilde{K} = K\sqrt{1-e^2}$ [m/s]	12.65(13)	3.20(17)	1.69(13)
e	0.022(10)	0.040(44)	0.195(73)
ω [°]	38(25)	122(61)	38(24)
λ [°]	142.89(64)	102.5(3.4)	62.2(4.6)
$M\sin i$ [$M_\oplus$]	15.86(16)	5.38(28)	1.77(13)
a [AU]	0.04061189(34)	0.0729286(80)	0.02845621(95)

	HARPS	Keck
data series and common fit parameters		
c_0 [m/s]	−9206.01(28)	0.86(34)
σ_{white} [m/s]	0.33(40)	1.18(28)
σ_{red} [m/s]	2.05(19)	
τ_{red} [day]	11.0(3.4)	
r.m.s. [m/s]	2.43	2.96
$\tilde{l}$ [m/s]	2.10	

Same notes as in Table 1.

necessary regularity of the RV data is lost at large time separations). Then why this red noise cannot be solely induced by some extra planets? Indeed, when working only in the time domain (correlation functions) it may appear practically impossible to disentangle the red noise from deterministic long- and moderate-term variations. However, in our work we rely on the frequency domain (power spectra), where this task is not that hard.

Of course, it remains theoretically possibile that this red noise represents just a mixture of many *true* periodic variations. Actually, the same logic can be applied to the usual white noise equally well. For practice, this interpreta-

tion really changes nothing: we still unable to model these hypothetical variations separately and we have to find some compound model for them. It is the case where we should just apply the Occam razor. The really meaningful question is: can we efficiently eleminate the red noise by means of only a few periodic harmonics? We find that in the case of GJ581 the answer is no. We started to honestly select the highest peaks from the last white-noise residual periodogram in Fig. 8 and subsequently remove the relevant periodicities from the residuals, one by one, each time plotting a new residual periodogram. We stopped after two such extra periodicities, because we did not achieve any impres-

Table 3. Best fitting parameters of the GJ581 planetary system: shared red jitter, four planets

	planetary parameters			
	planet b	planet c	planet d	planet e
P [day]	5.368603(66)	12.9198(20)	66.56(12)	3.14905(15)
$\tilde{K} = K\sqrt{1-e^2}$ [m/s]	12.62(13)	3.18(16)	1.81(28)	1.73(13)
e	0.022(10)	0.039(43)	0.28(11)	0.167(71)
ω [°]	32(26)	138(62)	329(26)	41(26)
λ [°]	143.06(63)	105.4(3.2)	143.0(8.8)	62.5(4.5)
$M\sin i$ [$M_\oplus$]	15.83(16)	5.35(26)	5.247(80)	1.81(13)
a [AU]	0.04061196(33)	0.0729331(75)	0.21754(25)	0.02845622(93)
	data series and common fit parameters			
	HARPS	Keck		
c_0 [m/s]	−9205.96(22)	0.79(31)		
σ_{white} [m/s]	0.50(27)	1.24(26)		
σ_{red} [m/s]	1.65(16)			
τ_{red} [day]	9.3(3.1)			
r.m.s. [m/s]	1.99	2.79		
$\tilde{l}$ [m/s]	2.01			

Same notes as in Table 1.

sive progress (the power excesses for periods more than 10 days and around the diurnal period were still their), and the residual periodogram contained already *three* moderate peaks at the periods of a few tens of days. They had the same marginal formal significance as the two previous ones. Clearly, only the red-noise models allow to purge out all these chameleonic peaks at once.

Based on our investigation, we conclude that the putative planets GJ581 *f* and GJ581 *g* likely do not exist, and the relevant RV signatures belong to the correlated noise. We do not say that we reject the existence of absolutely any planet in this system beyond the four known ones. However, even if some more planets exist in this system, they are not detectable in the present data, and they are unrelated to the peaks that we can see in the periodograms plotted with the use of only the white-noise model. The current RV data really support existence of no more than four planets orbiting GJ581.

6 CONCLUSIONS AND DISCUSSION

Although this work was focused on the concrete exoplanetary system of GJ581, we believe that our results may have much more general meaning. The cases of GJ876 discussed in (Baluev 2011) and of GJ581 are not unique, and it seems that there are more examples of planet-hosting stars demonstrating clear signs of the correlated noise in their publicly available RV data. Actually, we believe that the red RV noise might be rather common phenomenon.

This imposes bad as well as good concequences. The bad thing is that we have to use more complicated and computationally slow methods of the analysis. Without that any analysis of such data cannot be reliable. Unfortuanately, the functional shape of the correlation function have not yet been investigated well, so we have to make some rather voluntaristic guesses about it. Also, we need to accumulate rather large RV time series before the noise correlation pa-

rameters become fittable. In the case of GJ581, for instance, the size of the HARPS dataset is large enough, while the Keck data (half of the HARPS data in number) are not so good in this concern.

The good thing is that the method of the red-noise modeling does not just suppress the *phantomic* RV variations together with anything else on its way; it is capable to reveal *true* variations that were hidden beyond the fog of correlated noise. This means that our approach allows to increase, basically, the effective precision of the RV measurements, at least in the short-period domain. This offers a way to partly overcome the barrier set by the intrinsic RV jitter of the star, at least for some stars and in certain frequency ranges. In particular, we belive that our method may decrease exoplanetary detection threshold for active and/or subgiant stars, where the RV jitter contribution dominates in the total error budget, making it impossible to obtain the RV precision of better than $10 - 100$ m/s.

It is not yet fully clear, what is the source of the RV noise correlation. It may be caused by some long-living spots or other details on the star's visible surface, or may be a result of aggregation of various instrumental effects unrelated to the star itself. Our statistical analysis cannot offer a definite answer to this question. We believe that in general both reasons may be responsible. However, in the particular cases of GJ581 and GJ876 the first interpretation seems more likely, because in both these cases the red RV noise was detected in *two* independent datasets (HARPS and Keck).

In view of this topic, we cannot leave aside a very recent report on the detection of a "hot Earth" orbiting α Cen B (Dumusque et al. 2012). This discovery would not be possible without careful elimination of various effects of stellar activity polluting the periodograms at low frequencies. Basically, this team also applied some method of red noise modeling, based on its correlation with a spectral activity index. Although this method is different from our approach (we just unable to use the approach of (Dumusque et al. 2012), because all what we have is the public radial velocities of

GJ581 and not its raw spectra), we may note that the situation with the planet of α Cen B looks quite similar to the case of GJ581 *e*. Anyway, the work by Dumusque et al. (2012) puts more emphasis of our conclusion that the future of the Doppler planet searches lies in the careful treatment of the measured RV noise.

What concerns the particular case of the GJ581 planetary system, we were able to obtain several important results. First, we have shown that its RV data do not really support existence of any extra planet beyond the four-planet model. All apparent periodicities in these data, that were previously interpreted as extra planets, are illusions caused by RV noise correlations. Moreover, even the planet GJ581 *d* becomes doubtful. Its significance in the HARPS data does not even reach 2σ, and in the Keck data it is not detectable at all. In the combined dataset it may reach a more honourable level of 2.1σ but this level is still model-dependent. We admit that this planet is not rejected in our work and still remains rather probable, but despite of this fact we insist that it should be reclassified as a controversial one, until more data (preferably independent on HARPS) confirm it. The two-sigma significance level is not enough to claim a robust planet detection, if it was not confirmed by independent observations. On contrary, we were able to robustly confirm the planet GJ581 *e*. This planet can be revealed in the HARPS and Keck RV data independently, although to find its signal in the Keck time series it is mandatory to use a red-noise data model. The confirmation significance of this planet in the Keck data is $\sim 3\sigma$, although the same detection significance is only $\sim 1\sigma$.

At last, we would like to discuss briefly the results concerning the GJ581 planetary system presented by Tuomi & Jenkins (2012) in a recent preprint emerged during refereeing of our paper. In this work, the authors also used an autocorrelated noise model with exponentially decaying correlation, similarly to our work. Their main conclusion differing from ours is that using the Bayesian approach they can confirm the existence of the fourth planet, *d*, with larger statistical evidence. We must admit that this did not dissolve our criticism concerning this planet. We are of the opinion that such a difference between our conclusions was induced by different subjective prior distributions adopted in (Tuomi & Jenkins 2012), rather than on the objective information hidden in the RV data. Tuomi & Jenkins (2012) assumed the prior p.d.f. for planetary orbital periods as $\propto 1/P$, meaning a flat distribution in $\ln P$. It is easy to show that the periodogram approach used in our paper (which basically belongs to the family of AMLET tools) implicitly assumes a prior which is roughly flat in the *frequency*, implying the period distribution law $\propto 1/P^2$. Clearly, even if we leave behind the scene the discussion of what prior is better, the first prior dramatically shifts any data analysis in favour of longer-period signals, so there is no surprise that Tuomi & Jenkins (2012) obtained more significance for GJ581 *d* than we did. This difference is mostly subjective, however: we could reach roughly the same effect simply by dealing with renormalized periodograms, multiplying them by the value of the period in the abscissa.

What concerns the question which prior is better, the answer depends on the purpose of the analysis and other conditions. If our goal was to analyse a large array of datasets for many targets and to maximize the outcome of

this massive analysis, then we would use the prior $1/P$, because we know that the period distribution of real exoplanets is much more uniform in the log-period/log-frequency scale than in the linear-frequency one. However, here we deal with a planetary system that has high individual importance. For an individual dataset, the prior $1/P$ may induce some uncomfortable side effects. In particular, it favours to selection of longer-period aliases instead of the true periods. A practical example is provided by the two-planet system of HD208487. The Bayesian analysis assigns a 909 day period for the second planet of this system (Gregory 2007), but in the usual periodogram of the residuals there are actually *two* main peaks at 27 days and ~ 1000 days (Wright et al. 2007). These two peaks can be interpreted as monthly aliases of each other, and offer almost the same best-fit residuals r.m.s. However, the $1/P$ prior used in (Gregory 2007) just suppressed the shorter-period peak, allowing it to slip away from the view. In the case of GJ581, we must be even more careful with any pumping of long-period signals, because in such a manner we can easily pump up some red noise variation that our model was somehow unable to efficiently eliminate.

Anyway, returning to the planet GJ581 *d*, it is clear that its significance is not very impressive with the present RV data and, in addition, this significance is highly model-dependent. Under such circumstances, we prefer to follow the general principle of the frequentist approach in the statistics, which means that the selection between all realistic solutions must be done on the worst-case basis. GJ581 *d* needs to be confirmed by some independent non-HARPS data.

ACKNOWLEDGMENTS

REFERENCES

Baluev R. V., 2008a, MNRAS, 385, 1279
Baluev R. V., 2008b, Celest. Mech. Dyn. Astron., 102, 297
Baluev R. V., 2009, MNRAS, 393, 969
Baluev R. V., 2011, Celest. Mech. Dyn. Astron., 111, 235
Bard Y., 1974, Nonlinear Parameter Estimation. Academic Press, New York
Bonfils X., Forveille T., Delfosse X., Udry S., Mayor M., Perrier C., Bouchy F., Pepe F., Queloz D., Bertaux J.-L., 2005, A&A, 443, L15
Dumusque X., Pepe F., Lovis C., Ségransan D., Sahlmann J., Benz W., Bouchy F., Mayor M., Queloz D., Santos N., Udry S., 2012, Nature, 491, 207
Forveille T., Bonfils X., Delfosse X., Alonso R., Udry S., Bouchy F., Gillon M., Lovis C., Neves V., Mayor M., Pepe F., Queloz D., Santos N. C., Segransan D., Almenara J. M., Deeg H., Rabus M., 2011, arXiv:1109.2505
Gregory P. C., 2007, MNRAS, 374, 1321

Gregory P. C., 2011, MNRAS, 415, 2523

Lehman E. L., 1959, Testing Statistical Hypotheses. Wiley, New York

Marcy G. W., Butler R. P., Vogt S. S., Fischer D. A., Henry G. W., Laughlin G., Wright J. T., Johnson J. A., 2005, ApJ, 619, 570

Mayor M., Bonfils X., Forveille T., Delfosse X., Udry S., Bertaux J.-L., Beust H., Bouchy F., Lovis C., Pepe F., Perrier C., Queloz D., Santos N. C., 2009, A&A, 507, 487

O'Toole S. J., Tinney C. G., Jones H. R. A., 2008, MNRAS, 386, 516

Pont F., Zucker S., Queloz D., 2006, MNRAS, 373, 231

Tadeu dos Santos M., Silva G. G., Ferraz-Mello S., Michtchenko T. A., 2012, Celest. Mech. Dyn. Astron., 113, 49

Tuomi M., Jenkins J. S., 2012, arXiv.org, 1211.1280

Udry S., Bonfils X., Delfosse X., Forveille T., Mayor M., Perrier C., Bouchy F., Lovis C., Pepe F., Queloz D., Bertaux J.-L., 2007, A&A, 469, L43

Vogt S. S., Butler R. P., Haghighipour N., 2012, Astron. Nachr., 333, 561

Vogt S. S., Butler R. P., Rivera E. J., Haghighipour N., Henry G. W., Williamson M. H., 2010, ApJ, 723, 954

Wright J. T., Marcy G. W., Fischer D. A., Butler R. P., Vogt S. S., Tinney C. G., Jones H. R. A., Carter B. D., Johnson J. A., McCarthy C., Apps K., 2007, ApJ, 657, 533

APPENDIX A: SOME TIPS ON THE MAXIMUM-LIKELIHOOD ALGORITHM

A1 On the inverse of the noise covariance matrix

Possibly the fastest way to invert a real symmetric positive-definite matrix like $\mathbf{V}$ is to use the famous Cholesky decomposition: $\mathbf{V} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower-triangular matrix. It requires approximately $N^3/6$ floating-point multiplications. Moreover, having the matrix $\mathbf{L}$ at our disposal, we usually do not need to evaluate the inverse $\mathbf{V}^{-1}$ at all, because in true we usually need to evaluate only the matrix-vector combinations like $\mathbf{L}^{-1}\boldsymbol{x}$ or $\boldsymbol{x}^{\mathrm{T}}\mathbf{L}^{-1}$, which obviously can be obtained using the forward or back substitution. Moreover, these operations require practically the same CPU time as the direct multiplication by the precalculated inverse $\mathbf{V}^{-1}$ would do.

However, there is a single occurrence where the evaluation via direct matrix inversion seems the fastest way possible. It is the expression $\mathrm{Tr}(\mathbf{V}^{-1}\partial\mathbf{V}/\partial p_i)$ in (3). It seems that this task requires $\sim N^3$ floating-point operations (FLOPs) anyway. However, since this expression must be evaluated for many parameters p_i, it is faster to precalculate $\mathbf{V}^{-1}$ based on the Cholesky decomposition (this inversion requires $N^3/3$ floating-point multiplications) and then to evaluate the necessary matrix trace directly. Notice that the trace of a matrix product $\mathrm{Tr}\,\mathbf{AB}$ is equal just to the scalar product of the matrices involved, $\sum_{i,j} A_{ij}B_{ij}$, and thus requires only $\sim N^2$ operations.

A2 Avoiding matrix multiplications

Let us consider the calculation of $F_{p_i p_j}$ in (7). Even if we have precalculated the inverse $\mathbf{V}^{-1}$ or use some decomposi-

tion of $\mathbf{V}$ that makes its inversion easy, the expression (7) involves a few matrix multiplications of very large ($N \times N$) matrices, which require $\mathcal{O}(N^3)$ FLOPs. This is unsatisfactory and motivates us to find another representation for $F_{p_i p_j}$ that could be evaluated more quickly. Using the general identity $\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} = \mathrm{Tr}(\mathbf{A}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ and the relation $\mathbb{E}(\boldsymbol{r}\boldsymbol{r}^{\mathrm{T}}) = \mathbf{V}$ (the equality is exact because we should take the mathematical expectation at the true values of the parameters), we can transform the first of the expressions (4) as follows:

$$
\begin{aligned}
F_{p_i p_j} &= -\mathbb{E}\frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j} = \\
&= -\frac{1}{2}\mathrm{Tr}\left[\mathbf{V}^{-1}\left(\frac{\partial\mathbf{V}}{\partial p_i}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_j} - \frac{\partial^2\mathbf{V}}{\partial p_i \partial p_j}\right)\right] + \\
&\quad + \mathbb{E}\left[\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\left(\frac{\partial\mathbf{V}}{\partial p_i}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_j} - \frac{1}{2}\frac{\partial^2\mathbf{V}}{\partial p_i \partial p_j}\right)\mathbf{V}^{-1}\boldsymbol{r}\right] = \\
&= -\frac{1}{2}\mathrm{Tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_i}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_j}\right) + \\
&\quad + \mathbb{E}\left(\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_i}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_j}\mathbf{V}^{-1}\boldsymbol{r}\right) \quad\quad (A1)
\end{aligned}
$$

Performing the same transform leading to a matrix trace once again, we obtain the final expression for $F_{p_i p_j}$ in (7). Now, what if we apply this last transform in the opposite direction? Then we obtain the following approximation:

$$
F_{p_i p_j} \simeq \frac{1}{2}\boldsymbol{r}^{\mathrm{T}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_i}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial p_j}\mathbf{V}^{-1}\boldsymbol{r}, \quad\quad (A2)
$$

which has a relative error of the order of $1/\sqrt{N}$ (appearing because we also removed the expectation operator).

Since we use the Fisher matrix just as a handy approximation of the Hessian with the same relative error of $\mathcal{O}(1/\sqrt{N})$, the approximation in (A2) is no worse. However, (A2) can be evaluated without use of matrix multiplications. Already having the Cholesky decomposition $\mathbf{V} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, we can easily perform the first matrix-vector multiplication $\mathbf{V}^{-1}\boldsymbol{r} = (\mathbf{L}^{-1})^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{r}$ (only $\sim N^2$ FLOPs). After that we need to perform yet a few matrix-vector multiplications to form $\partial\mathbf{V}/\partial p_i \mathbf{V}^{-1}\boldsymbol{r}$ for all i (also $\sim N^2$ FLOPs). Then we need to multiply these vectors by $(\mathbf{L}^{-1})^{\mathrm{T}}$ from the left side (again $\sim N^2$ FLOPs) and evaluate the pairwise scalar products of the resulting vectors to obtain $F_{p_i p_j}$ for all i and j (only $\sim N$ FLOPs). This optimized procedure requires only $\mathcal{O}(N^2)$ FLOPs instead of the original $\mathcal{O}(N^3)$ FLOPs.

A3 Profit from the matrices sparseness

Since the noise correlation timescale that we are dealing with is about 10 days, while the total time span has the order of 10^3 days, most elements in the matrix $\mathbf{V}$ are close to zero. Therefore it is highly desirable to set small off-diagonal elements to zero exactly, and apply some algorithm of Cholesky decomposition and/or inversion tuned for sparse matrices. It is important that the first thing must be done in a smooth manner: we cannot just abruptly set all correlations below some small level to zero, since for the sake of smooth work of our Levenberg-Marquardt algorithm, we need to have continuously varying gradient and Hessian of the likelihood function. We reach this goal by means of the following smooth replacement in the argument of the noise correlation function: $\rho'(x) = \rho(x'(x))$, where $x' = x/(1 - (x/x_0)^2)$ and

x_0 is such that $\rho(x_0)$ is equal to some small value, e.g. 0.01. For $x > x_0$ we set $\rho'(x) \equiv 0$. After such modification, most of the elements in **V** become exact zeros. Interestingly, after that we noted a remarkable speed-up of various linear algebra calculations, even with no use of any specialized sparse-matrix algorithms. This indicates, probably, that modern CPUs execute various floating-point commands faster when one of the arguments is zero. With the use of algorithms tuned for sparse matrices, the performance increases even more dramatically. We unfortunately cannot give any reference or recommendation of any relevant software package, since the algorithms that we used in this paper we programmed ourself.

APPENDIX B: MONTE CARLO SIMULATION SCHEMES USED IN THE PAPER

B1   Plain Monte Carlo assuming Gaussian noise

(i) First of all, select some mock "true" values of the model parameters somewhere in the region of interest. We may select, for example, the nominal ones given in Table 1 for the white-noise model or analogous best fitting values for the red-noise models, although such choice is not mandatory.

(ii) Given the chosen vector of "true" parameters, evaluate the "true" RV values and the compound RV errors variances (and also correlations for red-noise models).

(iii) Construct a simulated RV dataset by means of adding to the evaluated RV curve the simulated Gaussian errors, generated on the basis of previously evaluated uncertainties and correlations.

(iv) Based on simulated dataset, evaluate the value of the likelihood function at the true parameter values from step 1, and the maximum value of the likelihood function for this trial. Based on these two values, evaluate the modified likelihood ratio statistic $\tilde{Z}$ for this trial, which is defined in (Baluev 2009).

(v) Save the newly generated value of $\tilde{Z}$, as well as the set of the simulated best fitting parameters (when necessary), and return to step 3, if the desired number of trials has not been accumulated yet.

B2   Bootstrap simulation

(i) Evaluate the best fitting model and the resulting RV residual.

(ii) Apply random shuffling procedure separately to the HARPS and Keck sets of the residuals.

(iii) Evaluate the statistic $\tilde{Z}$ and best fitting parameters in the same manner as in the plain Monte Carlo simulation.

(iv) Save the resulting value of $\tilde{Z}$ and parameters and return to step 2.

Note that the boostrap simulation is meaningful only when it is used with a white-noise RV model, because random shuffling of the residuals basically destroys any correlational structure of the RV noise, which a red-noise model tries to deal with.

B3   Genuinely frequentist Monte Carlo simulation

(i) Select an ith trial point in the space of model parameters $\boldsymbol{\xi}$ (or residing inside some given parametric domain).

(ii) Run the algorithm B1 assuming that true parameters correspond to the selected point.

(iii) Save the simulated distribution $P_i(\tilde{Z})$ of the test statistic of interest ($\tilde{Z}$ in our case) and return to step 1.

(iv) When a sufficiently dense coverage of the mentioned in step 1 parametric domain is reached, evaluate the function $P(\tilde{Z}) = \min P_i(\tilde{Z})$.

After that, the rigorous frequentist false alarm probability associated with an *observed* value $\tilde{Z}_*$ (that was obtained using exactly the same models that were used during the simulation) can be calculated as $1 - P(\tilde{Z}_*)$. This is a worst-case assumption method, in other words. Note that if we would stand on the Bayesian ground, we would evaluate, basically, some weighted average of $P_i(\tilde{Z})$ instead of the minimum, and this would force us to assume some prior distribution of the parameters. Obviously, in the frequentist approach we need only to circle a parametric domain, since any prior density inside this domain does not play any role when we find the minimum.

APPENDIX C: AMLET APPLICABILITY TO THE GJ581 CASE

Let us first freshen in brief a few practical things that AMLET includes:

(i) The asymptotically unbiased estimations of the model parameters are provided by the position of the maximum of the likelihood function.

(ii) Asymptotically, these estimations follow the multivariate Gaussian distribution with the covariance matrix expressed (again asymptotically) as the inverse of the Fisher information matrix, defined in (6).

(iii) To test some simple "null" model against a more complicated alternative one (which encompasses the null hypothesis as a partial case), we need to construct the relevant likelihood ratio statistic, and evaluate the false alarm probability associated with the null hypothesis rejection. The latter false alarm probability can be found from the known asymptotic χ^2 distribution of the likelihood ratio logarithm.

(iv) Consequently from the previous point, the multi-dimensional confidence regions for some set of model parameters are outlined as level curves (or level surfaces) of the likelihood function, considering it after maximization over the rest of free parameters. The value of the likelihood ratio corresponding to the global maximum and a given level curve yields the confidence probability of this level curve (again assuming the asymptotic χ^2 distribution for logarithm of this ratio).

When the fit model is linear or well linearisable, AMLET is accurate already for relatively small number of observations. When the model non-linearity increase, the critical number of observations, after which AMLET becomes applicable, appears impractically large, so that for practical numbers AMLET offers poor precision. Since in our case of GJ581 we deal with rather complicated non-linear model of the RV data, we would like to find out, which AMLET

proposition we can be used safely under our concrete circumstances?

We may notice that the AMLET proposition listed above demonstrate different resistance with respect to a change of variables (re-parametrization of the original model). For example, assume we have some model parameter x, and we make a replacement $x \longmapsto y = 1/x$, treating the old primary fit parameter x as only a derived one. Even if the distribution of the estimation of x was exactly Gaussian, the analogous distribution of y may appear completely non-Gaussian. Therefore, rather formal action like a non-linear change of variables, which did not really alter the functional structure of our original model, was able to invalidate some of the AMLET propositions. However, some other propositions, namely those dealing with only maxima of the likelihood function, remained intact. Indeed, the maximum value of a function is invariable with respect to any change of the independent variables (at least if this change is a one-to-one mapping), so the quantities like the likelihood ratio statistic are invariable with respect to such re-parametrization. This phenomenon is called sometimes as exogenous and endogenous non-linearity. The exogenous non-linearity does not belong to the physics of the original task, and depends on human-controllable things like, for instance, the choice of the system of free parameters, time reference point, coordinate system, etc. The endogenous non-linearity represents an immanent property of the task and it cannot be eliminated by any such trick.

Endogenous part of the non-linearity is the only thing that we really need to take care of, since anything beyond it is, basically, just a result of incarefully chosen parametrization. Since in this paper we mainly deal with the likelihood ratio test and its descendants, we need to check how precisely the asymptotic χ^2 distribution can approximate the real distribution of the relevant likelihood ratio statistic. We consider three models to verify: the 4-planet white-noise model, 4-planet shared red-noise model, and 4-planet separated red-noise model. For each of these models, we perform the Monte Carlo simulation sequence B1 of the Appendix B.

For the white-noise model, we also perform two bootstrap simulation series, which is a popular tool for exoplanetary RV data analysis works (e.g. Marcy et al. 2005) due to its resistance to possible non-Gaussian errors in the data. The first simulation is done according to the scheme B2 in the Appendix B, while in the second bootstrap simulation we applied the same algorithm to a simulated dataset with purely Gaussian noise (rather than to the real RV dataset).

The results of these simulations are shown in Fig. C1. For the white-noise model, we find that the simulated distribution of $\tilde{Z}$ is in amazingly perfect agreement with the asymptotic χ^2 approximation. For the shared red-noise model the agreement is still good. And only for the separated red-noise model AMLET offers poor precision.

It is rather unexpected that in the white-noise case bootstrap simulation disagrees both with the Monte Carlo and with the χ^2 distribution. This disagreement does not indicate that the RV noise in the real data is non-Gaussian. Vice versa, two bootstrap curves lie very close to each other, meaning that real RV measurement errors are indistinguishable from Gaussian noise. This means that the bootstrap itself is basically an inadequate tool for our tasks.

So why bootstrap did not work in Fig. C1 as we expected? We find that the reason is contained in the RV jitter parameters. Investigating Fig. C2, we can see that while the usual Monte Carlo simulations are again in good agreement with AMLET confidence contours for the HARPS and Keck jitter, the results of the bootstrap are definitely wrong: they are biased and locked in an inadequately narrow region of the plane. The reason of this behaviour is clear: while we shuffle the best-fit residuals, their scatter remains constant, and since the RV jitter is derived mainly from this scatter, such shuffling keeps both jitter estimations almost constant. This means that bootstrap cannot be applied to data models involving some parameters of the noise.

Now let us investigate the non-linearity of our RV models in a bit more depth. We consider 2D confidence regions for a few pairs of model parameters. For this goal, we select three pairs of parameters that demonstrate largest mutual correlations, since such parameters are usually affected by stronger non-linearity effects (Baluev 2008b). As it turned out, all these pairs involve the mean longitude λ and the pericenter argument ω of one of the planets. The relevant asymptotic 2D confidence contours, constructed on the basis of our statistic $\tilde{Z}$, are shown in Fig. C3. Clearly, they have little common with ellipses, that we would see for a well-linearisable model. However, this non-linearity has only exogeneous nature and is caused, obviously, by small planetary eccentricities, which make the parameter ω poorly determinable. Under such circumstance we should better consider, instead of the pairs (λ, ω) the pairs $(\lambda, e\cos\omega)$ and $(\lambda, e\sin\omega)$, since the parameters $e\cos\omega$ and $e\sin\omega$ are more adequate than e and ω, when dealing with almost circular orbits.

Our conclusion, that the apparent non-linearity of these parameters is only exogenous, is confirmed by numerical simulations, which are in good agreement with the asymptotic confidence contours. The agreement is equally good for the bootstrap and for the pure Monte Carlo methods, which generate practically identical sets of points. This also confirms our previous conclusion that the RV data for GJ581 do not show any detectable non-Gaussianity.

And the final question, why the separated red-noise model demonstrates so large deviation from AMLET's χ^2 distribution in Fig. C1? What is the source of this endogenous (and thus more important) non-linearity? Obviously, the reason is hidden in the RV noise model, because we have already established that the RV curve model may produce only negligible endogenous non-linearity. To investigate this question in more depth, we performed the same Monte Carlo simulation treating the HARPS and Keck data entirely independently. We again assumed the separated red-noise model for these datasets. What concerns the RV curve model, for the HARPS data we adopted the four-planet one, while for the Keck data we considered three-planet and two-planet models (with and without planet e, and with no planet d). After that we simulated the distribution of the statistic $\tilde{Z}$, and compared it with the relevant asymptotic χ^2 distribution, as in Fig. C1. We obtained that for the HARPS dataset the agreement is the same good as for the shared red-noise model, while the Keck dataset demonstrates bad things (Fig. C4).

Therefore, the main source making the separated red-noise model statistically poor, is the Keck dataset, which
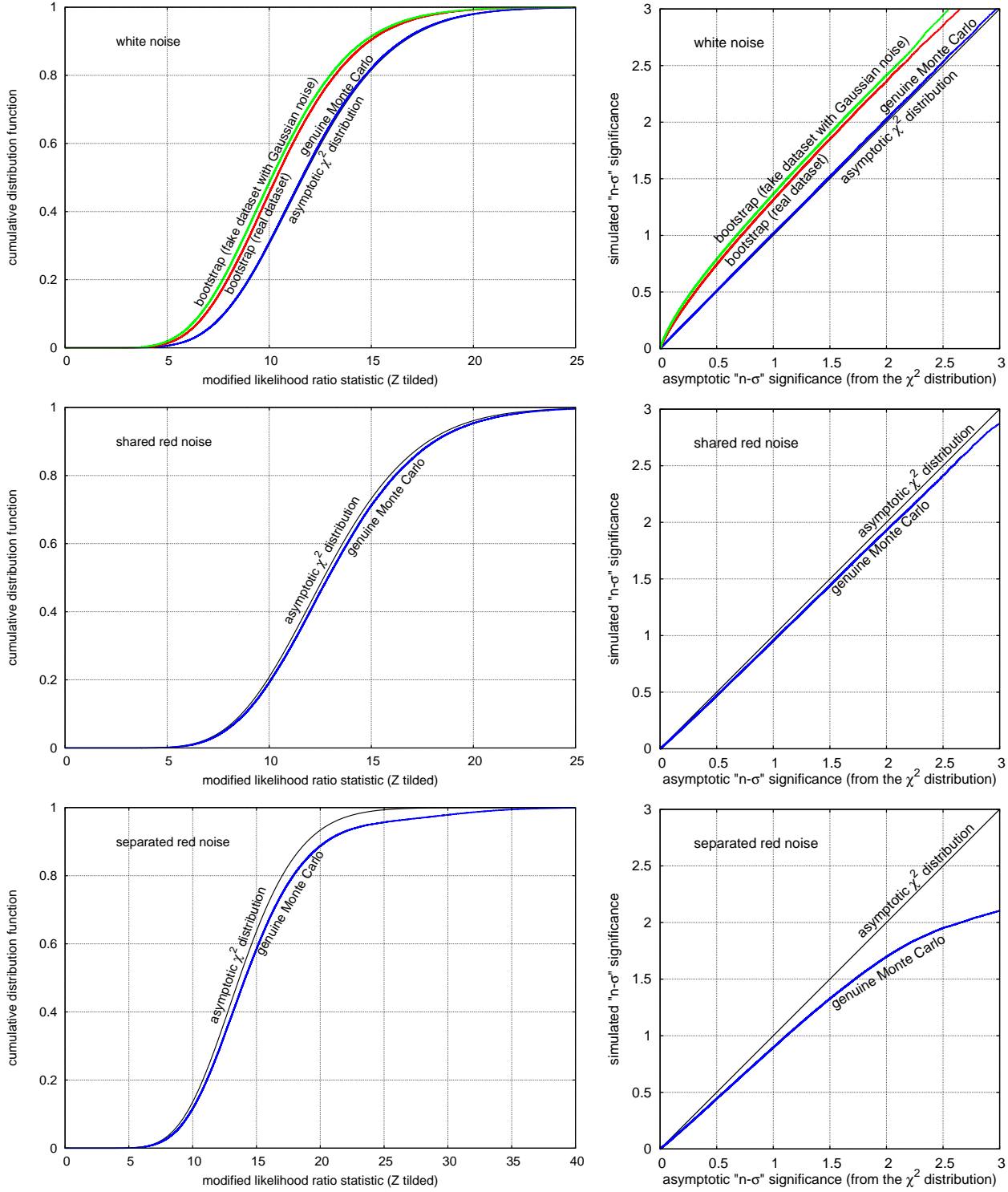
Figure C1. Simulated distributions of the statistic $\tilde{Z}$ from (Baluev 2009) and their asymptotic χ^2 approximation. The null hypothesis was that the model parameters are equal to the best fitting values (which are treated as true during the relevant simulation), and the alternative was that all these parameters are unknown (free). The plots in the left column show the simulated cumulative distributions as functions of $\tilde{Z}$, while the relevant simulated significance levels as functions of the asymptotic χ^2 significance are shown in the right column. The top, middle, and bottom pairs of panels differ by the assumed RV noise model, according to the marks in each graph. The RV curve model is always the four-planet one. Note that different noise models imply different number of degrees of freedom, so the χ^2 distributions become different too.

Likelihood confidence contours for estimated RV jitter of GJ581
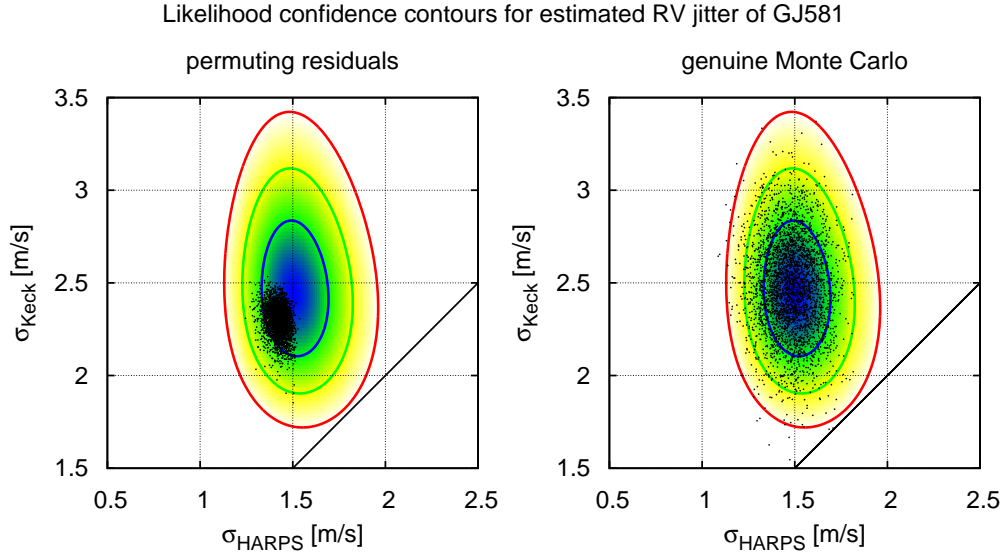

Figure C2. Same as Fig. C3, but for the RV jitter parameters of the HARPS and Keck data.

can provide only rather ill estimations of the RV noise parameters, when it is used without HARPS data.

This paper has been typeset from a TEX/ LATEX file prepared by the author.

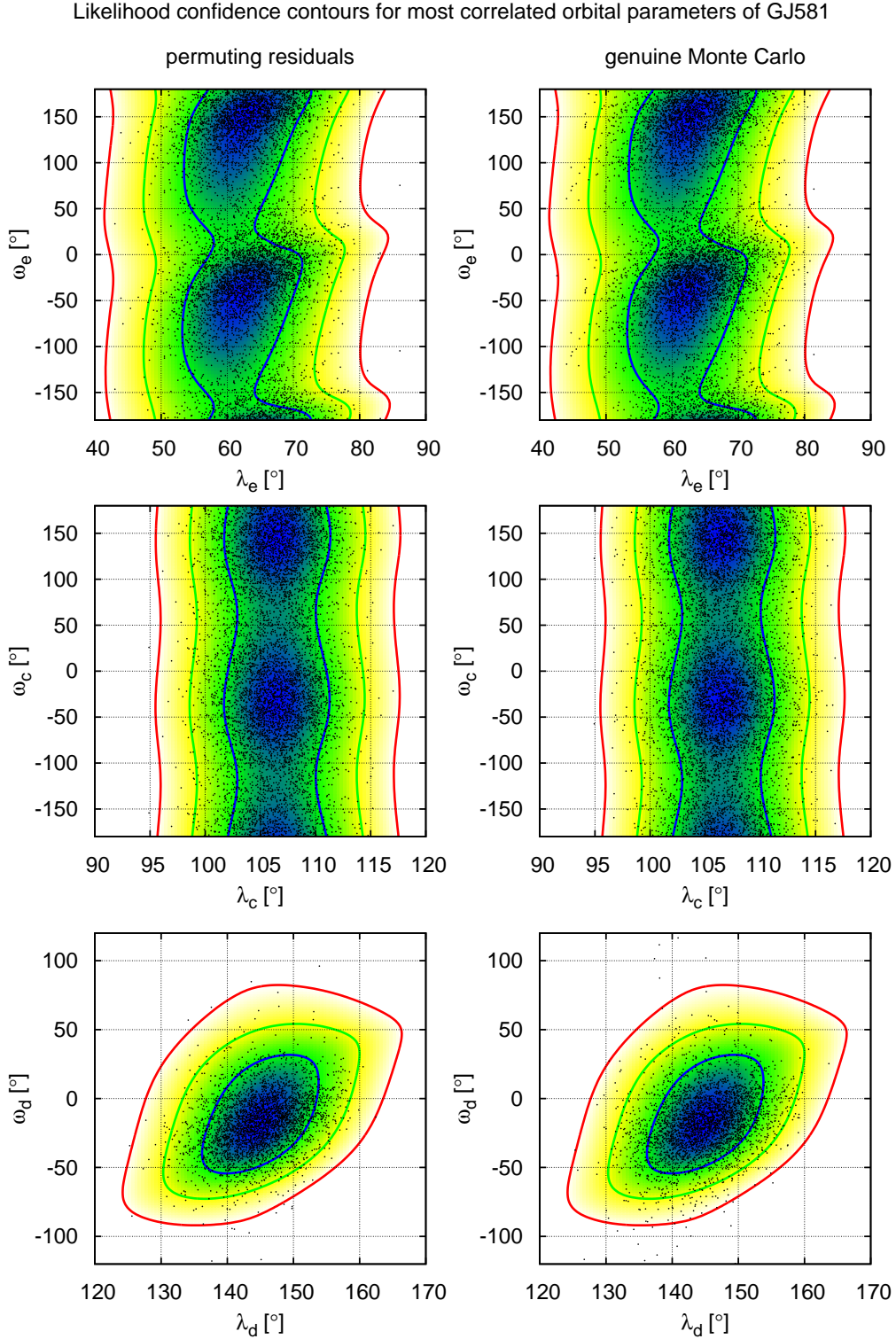Likelihood confidence contours for most correlated orbital parameters of GJ581

Figure C3. The asymptotic confidence contours for a few most correlated pairs of parameters of the GJ581 4-planet white-noise model, in comparison with Monte Carlo and bootstrap simulations. These asymptotic confidence contours represent the level curves of the modified likelihood ratio statistic $\tilde{Z}$ (Baluev 2009). These contours are shown by means of colormaps and three reference level curves corresponding to the asymptotic 1-, 2-, and 3-σ significance levels (derived using the asymptotic χ^2 distribution of $\tilde{Z}$). These contours are identical for the plots in the left and right columns; the things that differ are the simulated points, that were obtained by means of the bootstrap (permuting best-fit RV residuals) or pure Monte Carlo (generating Gaussian noise based on the best-fit model). The number of simulated points shown in each graph is 3333.
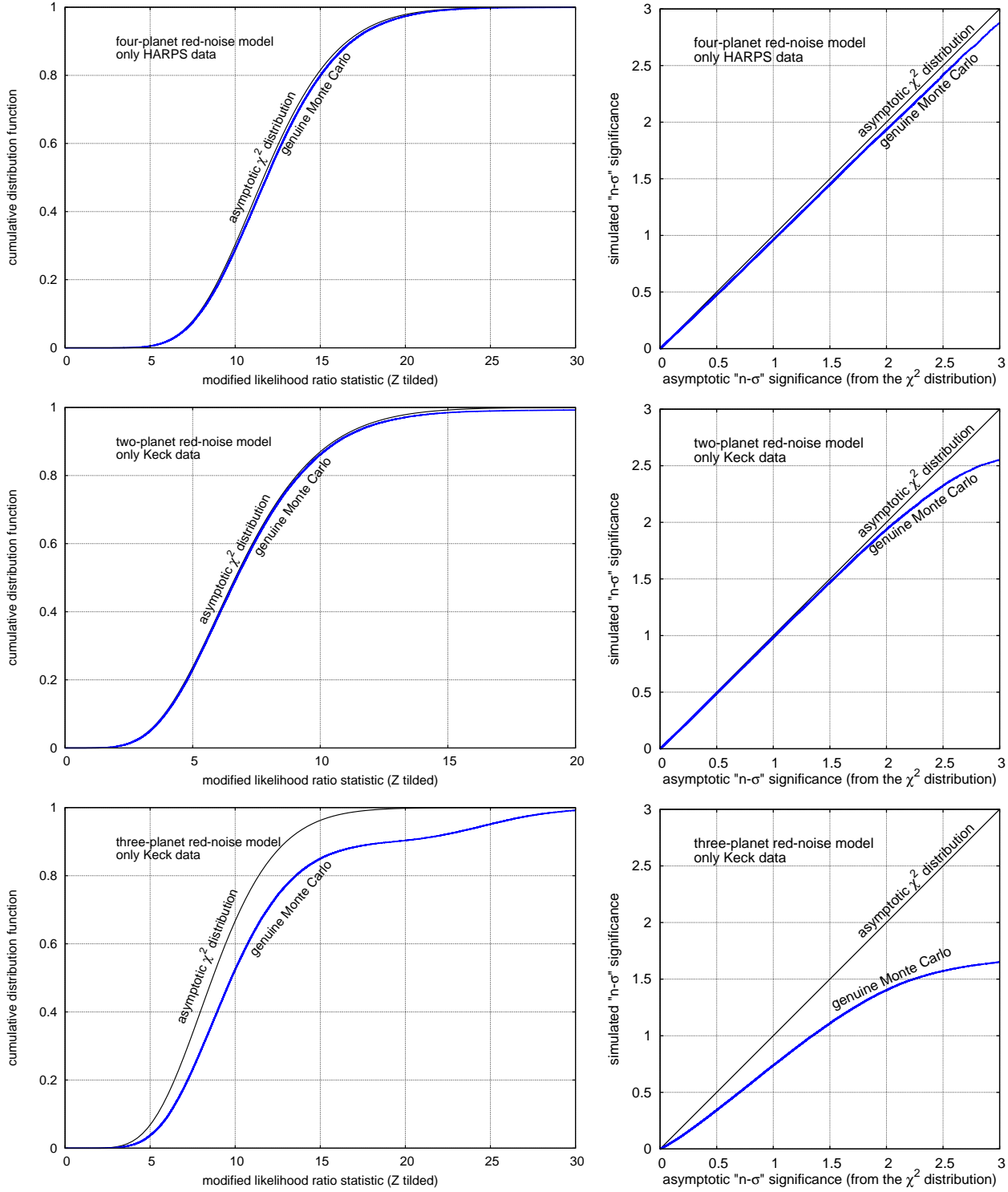
Figure C4. Same as Fig. C1, but fitting the HARPS and Keck time series entirely independently from each other. The top panels correspond to the HARPS dataset with four-planet model; middle pair – to the Keck dataset with two-planet (*b* and *c*) model; bottom pair – to the Keck dataset with three-planet (*b*, *c*, *e*) model. The noise model in each case always incorporated the red component.